Revaluating Semantometrics from Computer Science Publications

Christin Katharina Kreutz, Premtim Sahitaj, and Ralf Schenkel

Trier University, 54286 Trier, DE
{kreutzch, sahitaj, schenkel}@uni-trier.de

Abstract. The identification of important publications is subject in many research projects. While the influence of citations in finding seminal papers has been analysed thoroughly, semantic features of citation networks are regarded with less vigour. In this paper, we revaluate the ideas of semantometrics presented by Herrmannova et al.[9,13] to learn patterns of features extracted from publication distances in their citation networks aiming at distinguishing between seminal and survey papers in the area of computer science. For the evaluation, we present the SeminalSurveyDBLP dataset. By using different document content representations, the incorporation of semantic distance measures, as well as multiple machine learning algorithms for the classification, we achieved an accuracy of up to 0.8015 on our dataset. Earlier findings in this area suggest features extracted from references to be more suitable proxies whereas we observed the contrasting importance of features describing citation information.

Keywords: Semantometrics · Classification · Citation Network · Natural Language Processing.

1 Introduction

With the ever growing amount of scientific publications, automatic methods for finding influential or seminal works are indispensable. While the majority of research tackles the identification of important works [9,39,35,40,7,38], the influence of semantic features in this context has not been explored thoroughly. Citation based classification or impact measures are dataset dependent [30,20]. They need to be handled with care due to self-citations [29], varying citation practices in different areas [5,30,32], diverging reasons for citing [6], the non-existence of citations of new papers [38] and uncited influences [23,19,6].

Distinguishing between seminal publications which advance science and popular survey papers might pose a problem as both types are typically cited often [30] but reviews are over-represented amongst highly cited publications while not contributing any new content [1]. Seminal papers are ones which are key to a field while surveys review and compare multiple approaches and can be comprehensible summaries of a domain. Influential members of both classes can be distinguished from all other publications by observing their number of citations.

Differentiating between seminal and review papers is not as simple. Therefore, methods considering more factors than the number of citations and references are desirable [38,20] as these values are no sufficient proxy in measuring publication impact and scientific quality [11,30]. Preferably, an approach with the potential to measure the contribution of a paper and how much it advanced its field should be favoured.

Herrmannova et al. [9] assume the classification of a paper as seminal or survey can be performed by observing semantometrics as a new metric for research evaluation which uses differences in full texts of a citation network to determine the contribution or value of a publication [13]. They conducted their experiments on a multi disciplinary dataset [11]. We want to access the usefulness of this approach and revaluate the provided ground truth on a dataset restricted to a less broad area.

Our contribution is two-fold: We introduce SeminalSurveyDBLP, a dataset suitable for the task of classifying a publication with usage of its citations and references as seminal or survey paper. Additionally, we analyse the approach presented by Herrmannova et al. [9] using different document representations as well as numerous classification algorithms and evaluate the usage of single and multiple features in the classification process on the new dataset in a different and more homogenous domain.

The remaining content of this paper is organized as follows. Section 2 gives an overview of the already established conceptual background and related research. In Section 3, the SeminalSurveyDBLP dataset is presented. The succeeding Section 4 introduces utilized document vector representations, distance measures and classification algorithms to revaluate Herrmannova et al.'s approach [9] on the new dataset. A detailed evaluation of our dataset on different feature modes is given in Section 5.

2 Background and Related Work

2.1 Background

Extraction of mathematical descriptors from data is common in medical image analysis [8,15] but for publication networks, it was initially introduced as semantometrics in [13] to access research contribution.

Herrmannova et al.'s approach [9] which uses these principles described in [13] is the base of this revaluation. They were the first to work with citation distance networks (CDN) that each centre around a publication P which is connected to references X and citations Y to classify if P is a seminal or survey paper. Semantic distances that describe the relationships between publications were measured: The distances between titles and abstracts of X and Y are contained in group A, distances between a publication and its references are included in group B and group C is composed by distances between P and its ingoing citations. The semantic distances between entries of X can be found in group D, symmetrically, distances between citing publications are stored in E.

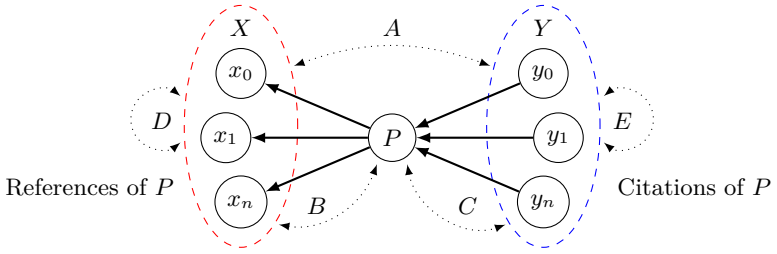

Figure 1: Neighbourhood of publication P. Nodes symbolize publications, edges between papers represent citations. $\{x_0, \ldots, x_n\}$ are references of P, $\{y_0, \ldots, y_n\}$ are citations of P. Dotted edges symbolize observed relationships between publications.

Figure 1 visualizes the different groups of relations in a CDN. From these five groups, 12 features each were extracted such as min, max and mean distance. A classification accuracy of 0.6897 was achieved by using naïve Bayes on single features derived from cosine distances between tf-idf vectors of papers' abstracts from the TrueImpactDataset [11]. This dataset was created from a user study and contains publications from multiple distinctly labelled fields.

2.2 Related Work

Relevant topics for our work besides semantometrics are text-based methods and prediction of influence using citation networks.

Several papers can be found in the field of *language-based methods and citation networks*. Topical developments of documents with identification of influential publications [7] or similarity of full texts of citations [39] are appropriate proxies in determining publication impact. Prediction of citation counts using text-similarity of extracted popular terms of publications [18] is also rooted in this domain. Context-aware citation analysis on full texts and leading edge impact assessment [23] and content similarity of abstracts of citing publications as well as the cited papers [37,40,26] are able to identify important references of publications.

For *prediction of influence based on the citation network* of a publication, a measure based on the desired audience or purpose of a paper [22], the fluctuation or stability of members in research teams (research endogamy) [21,33,10] as well as research contribution of individual authors measured on established links between communities [28] can be analysed.

3 SeminalSurveyDBLP Dataset

Half of the 1320 publications in our SeminalSurveyDBLP[31] dataset are seminal while the other half of papers are surveys. All works are from the area of computer science and adjacent fields as they are contained in dblp [17]. For seminal

statistics / class	# papers in class	avg. len abs of P	# references (X) #	# references (X) min	# references (X) max	# references (X) avg.	# citations (Y) #	# citations (Y) min	# citations (Y) max	# citations (Y) avg.
seminal	660	172.25	20,858	10	154	31.6	50,397	10	1370	76.4
survey	660	173.03	29,366	10	186	44.5	51,082	10	1365	77.4

Table 1: Numeric description of unstemmed SeminalSurveyDBLP dataset.

publications, entries published in conferences attributed as A* at the CORE Conference Ranking[4] such as *SIGIR*, *JCDL* or *SIGCOMM* were collected as publications often cited (and thus important) tend to appear in high-impact venues [1]. We assume papers published in a seminal venue as attributed by the CORE rank are seminal themselves, or they would not have been accepted for such a venue even if they have not yet accumulated large amounts of citations. This might be a strong assumption, as not every paper from an A* conference is seminal and seminal papers can also appear in other venues. Surveys were extracted from *ACM Computing Surveys*, *Synthesis Digital Library of Engineering and Computer Science* and *IEEE Communications Surveys and Tutorials*. These venues are specialized in solely publishing reviews. Every paper used has at least ten citations and references.

For each of the papers, the citations and references were collected. Citation information and abstracts from the AMiner dataset [36] were joined with dblp data to make sure they were also from computer science or adjacent domains. The join was based on matching DOIs of dblp papers with ones from AMiner or paper title and author name matches where DOIs were not present. Full texts are not included in the AMiner dataset. Citations and references not contained in dblp were omitted. For every paper, its year of release is also enclosed. Considered publications for P, X and Y needed to have a length of at least ten terms in their combined title and abstract. The dataset is engineered so that there are similar numbers of citations and references for publications of the opposing classes. The total number of unique publications contained in the dataset is 121,084. Table 1 shows statistics regarding the length of abstracts and number of citations for each type of paper for an unstemmed version of the dataset. As the increased amount of references is assumed to be a feature of survey papers compared to seminal publications, the average and total number of references is higher and thus our dataset is unbalanced in this aspect. Figure 2 shows the distribution of numbers of references and citations for all papers of groups seminal and survey from the dataset. Numbers of citations are distributed rather homogenous between the two classes, but for references, differences in the distributions can be seen. While there are fewer publications with a few references for surveys, a gap in the number of references from 40 to 50 can be seen for seminal papers.

Of the 660 seminal papers 24 have received a best paper award. Only incorporating publications which received an award would dramatically decrease the size of the dataset as a similar distribution of references and citations for works of both classes was prerequisite in its construction.

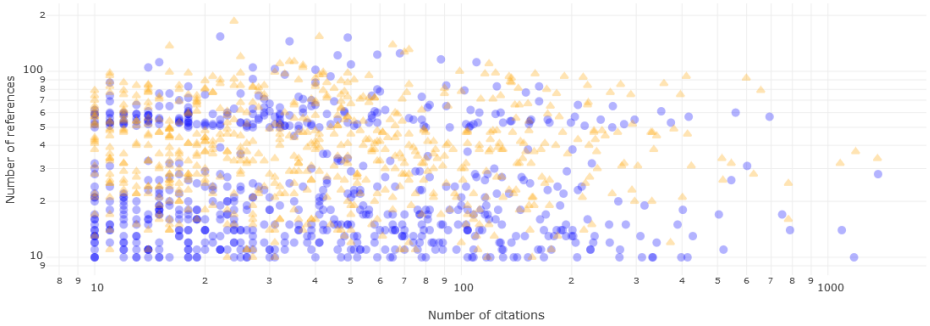

Figure 2: Distribution of number of references and citations for seminal (blue circles) and survey (orange triangles) publications.

4 Methodology

Herrmannova et al. [9] proposed the usage of citation distance networks to extract patterns from text which can be represented by distance features for making assumptions whether publications are seminal or survey. First, document representations of P, its references X and its citations Y need to be generated. In a next step, distances between publications for every group A to E can be calculated. From these sets of distances, 12 features are then computed each: Minimum, maximum, range, mean, sum of distances in a group, standard deviation, variance, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, skewness, and kurtosis. Those $12 \cdot 5 = 60$ features are named by concatenating the feature with the group it originates from like *minA* or *rangeE*. On these features, classification algorithms are able to predict the class seminal or survey a publication P should be associated with.

We extracted these features by using different configurations on which we will conduct our evaluation. For document vector representations (V) we used tf-idf to be able to compare our results with [9] directly as they also used this text representation and doc2vec [16] as possible improvement. The tf-idf values are computed on the 121,084 publications in the stemmed (S) or unstemmed (U) SeminalSurveyDBLP dataset, abstracts of all citations and references were included in the calculation of term frequencies. Weights for doc2vec (d2v) were generated using the English unstemmed Wikipedia corpus from 20th January 2019. We refrained from using doc2vec on a stemmed corpus as this preprocessing is no prerequisite for achieving good results [16]. The model was trained to consist of 300 dimensions with usage of distributed memory as learning algorithm.

As distance measures, cosine distance was applied as described in [9], additionally Jaccard distance was also used as a second method.

Classification algorithms (C) used are logistic regression (LR), random forests (RF), naïve Bayes (NB), support-vector machines (SVM), gradient boosting (GB), k-nearest neighbours (KNN) and stochastic gradient descent (SGD). In [9], SVM, LR, NB and decision trees were applied. We wanted to include those

V	tf-idf				d2v	
distance measure	Cosine		Jaccard		Cosine	Jaccard
	U	S	U	S	U	U
# significant features	27	31	30	32	50	27
overlap with [9] in %	48.48	54.55	54.55	54-55	78.78	39.39

Table 2: Number of significant features per document representation and distance metric. Overlap with significant features from [9] is rounded on two decimal places.

classifiers except for decision trees which we omitted as we incorporated random forests.

To find influential features for every combination of document vector representation and distance measure independent two-sided t-tests with p=0.1 were conducted. Table 2 shows the number of significant features for the different variants as well as the percentage of overlap when compared with the 33 significant features computed in [9]. Usage of doc2vec resulted in the highest number of significant features when combined with cosine distance. Here, the overlap in significant features from SeminalSurveyDBLP and TrueImpactDataset [9] is also the highest. In general, the intersection of significant features between the datasets is modest which indicates differences between them.

5 Evaluation

Python 3.7 and scikit-learn [24] implementations of classifiers were used in the evaluation process.

Based on the single significant features, all significant features, all features, the 33 significant features of the TrueImpactDataset and the features which were significant in the TrueImpactDataset as well as in the SeminalSurveyDBLP dataset were used in the classification process. All accuracies (A) and F1 scores (F1) are rounded to four decimal places. Values have been calculated by usage of ten-fold cross validation.

5.1 Single Publications

In a first step, classification solely on the vector representations of publications P is tested without consideration of citations and references. This approach is highly successful in determining the class of P. Wile using the doc2vec representation of the TrueImpactDataset [11] an accuracy of 0.6109 and F1 score of 0.5903 can be reached in using logistic regression. Restricting the TrueImpactDataset on publications from the area of *Computer Science and Informatics* to establish comparability with SeminalSurveyDBLP results in 37 publications with abstracts available. Classifying on this part of the dataset in doc2vec document representation, results in an accuracy of 0.7838 and F1 score of 0.7895

dataset	SeminalSurveyDBLP						[11]		[11] only CS	
V	tf-idf				d2v		d2v		d2v	
	U		S		U		U		U	
C	A	F1	A	F1	A	F1	A	F1	A	F1
LR	.5	.0	.7205	.612	.8932	.8942	.6109	.5903	.5946	.5714
RF	.9159	.916	.9197	.9202	.8583	.8569	.5481	.5222	.5946	.5455
NB	.7462	.7468	.7409	.7467	.8235	.8255	.5397	.4954	.3784	.303
SVM	.5985	.6355	.6311	.6484	.9182	.9193	.59	.5421	.4595	.2308
GB	**.9288**	.9258	.928	.9247	.8508	.8529	.5439	.5198	.7838	.7895
KNN	.5561	.6324	.572	.2766	.7803	.8124	.5983	.4074	.4324	.2222
SGD	.6553	.7387	.6826	.7516	.8629	.8642	.59	.5812	.3784	.303

Table 3: Classification accuracy and F1 scores for different algorithms based on document vector representations of publications P from SeminalSurveyDBLP, TrueImpactDataset [11] and only computer science (CS) papers from the TrueImpactDataset.

when applying gradient boosting. Meaningful tf-idf vectors could not be created for the TrueImpactDataset as the abstracts of citations and references are not contained in it.

Using unstemmed tf-idf representations of the SeminalSurveyDBLP dataset with gradient boosting results in an accuracy of 0.9288 with corresponding F1 score of 0.9258. For this task, tf-idf is a better proxy than doc2vec as highly descriptive terms such as survey or review are encoded in single features in the feature vector. Table 3 shows all results in detail.

Remarkably, the dataset proposed by us is much more suitable for classification of publications as survey or seminal than the TrueImpactDataset. This might be owed to the creation process of the SeminalSurveyDBLP dataset as there are papers chosen for class survey which originate from journals specialized on surveys. It is unsurprising that they oftentimes contain this or similar keywords in their title or abstract. As there is no easy way other than resorting to such means in order to create a sufficiently large database automatically, this property of the dataset is nearly ineluctable. Conducting user studies to find seminal and survey publications leads to the same problems which occurred in [11]. Multiple submitted publication titles could not be matched to the real papers, the different research areas are not evenly represented in the data and human bias or misjudgement cannot be eliminated. Another possible explanation for the good performance of our dataset might be the focus on one area. All publications are from the wider field of computer science. Publications contained in [11] originate from all different disciplines where there might be a multiplicity of ambiguous descriptions for surveys. This assumption is supported by the considerably good performance of the TrueImpactDataset restricted on publications solely attributed to come from computer science.

measure	Cosine distance								
V	tf-idf						d2v		
	U			S			U		
C	F	A	F1	F	A	F1	F	A	F1
LR	minE	.6152	.6482	minE	.6053	.6552	50pD	.6697	.6813
RF	sumE	.6561	.647	sumE	.6742	.6702	sumE	.6379	.6368
NB	rangeE	.6083	.6551	rangeE	.5939	.6642	50pD	.6705	.701
SVM	sumC	.722	.6959	sumC	.7205	.6938	sumE	.7167	.6934
GB	sumE	.7371	.7274	sumE	.7379	.7293	sumE	.7318	.7226
KNN	sumE	.7159	.7017	sumE	.7258	.7127	sumE	.7045	.6986
SGD	sumE	.6386	.7054	sumE	.6379	.7114	avgD	.6636	.6641
measure	Jaccard distance								
V	tf-idf						d2v		
	U			S			U		
C	F	A	F1	F	A	F1	F	A	F1
LR	rangeE	.6045	.5515	rangeE	.6174	.5781	25pD	.6015	.6086
RF	sumE	.6803	.6754	sumE	.675	.6667	sumE	.6205	.6225
NB	skewE	.6	.581	rangeE	.603	.5313	25pD	.5894	.6323
SVM	sumC	.7311	.7019	sumC	.7288	.6992	sumE	.7091	.6863
GB	sumE	**.7432**	.7362	sumE	.7409	.7328	sumE	.7121	.7139
KNN	sumE	.7121	.6984	sumE	.7288	.7163	sumE	.6758	.6698
SGD	sumE	.6402	.7151	sumE	.6371	.7023	25pD	.5902	.4967

Table 4: Observation of distance measures cosine and Jaccard with different document vector representations. For every classification algorithm, the single feature (F) achieving highest accuracy and the corresponding F1 score are displayed.

5.2 Single Features

In an additional experiment, the whole citation distance network of publications is considered for the classification task based on a single feature. Each of the 60 features derived from the CDN of a publication is used on its own as input for the machine learning algorithms. Classifying on these sole features lead to an accuracy of up to 0.7432 with an F1 score of 0.7362, when using gradient boosting with tf-idf on unstemmed documents and Jaccard as distance measure which is +0.0535 compared to the top value from [9]. In the most descriptive features, sum and range of C and E were contained mostly for tf-idf text representations. For doc2vec text representations sumE and percentiles from group D were found to be good predictors. Features from group D can be interesting as they are already present at the submission of the publication when a paper has yet to gain citations. For all combinations of document representations and distance measures, classifying on sumE lead to the best results.

Herrmannova et al. [9] found features of groups B, C and D to operate well for this task while we observed contrasting behaviour. The difference in outcome might be explained by the observation of citation practice of high impact pub-

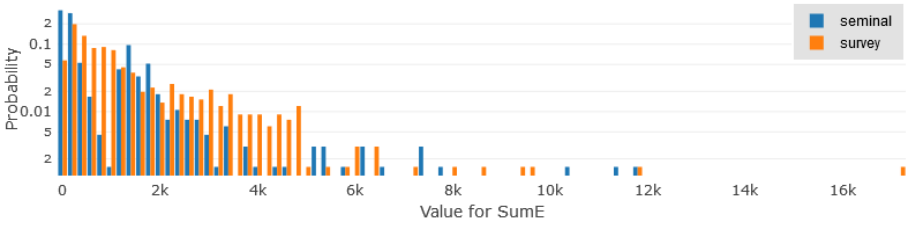

Figure 3: Probability distribution of feature achieving the highest accuracy.

lications. In computer science, they behave differently from those in other areas as they cite from a narrow community [32]. While our dataset is rooted in this one area, the TrueImpactDataset [11] spans multiple disciplines.

Generally, features extracted from doc2vec performed worse than the tf-idf ones. While gradient boosting is a good classifier for this task, logistic regression, naïve Bayes and stochastic gradient descent did in general not perform well. Table 4 shows the results of the single feature classification for every machine learning algorithm and document representation in detail.

Feature values are very similar distributed for cosine and Jaccard distance with the same document representation. Figure 3 shows the exemplary probability distributions for sumE, the best feature for unstemmed tfidf vectors combined with Jaccard similarity.

5.3 All Features

In the last experiments, the whole citation distance network of publications is again used for classification based on multiple features. Here, of the 60 features derived from the CDN of a publication, multiple features are used in combination as input for the machine learning algorithms.

Using cosine distance, tf-idf vectors perform worse than doc2vec document representations when considering all significant features for the classification task. When using Jaccard distance, tf-idf scores a higher accuracy than doc2vec. On our dataset Jaccard distance performs better with document representations of stemmed articles while no clear statement can be adhered for cosine distance. Overall, gradient boosting and random forests are the best machine learning algorithms in this context. The best accuracy of 0.8015 with corresponding F1 score of 0.8053 can be achieved using gradient boosting and cosine distance on doc2vec text representation. This is +0.0583 in accuracy when compared to the single feature variant.

When using all features to classify seminal and survey publications, a similar phenomenon with text representation and distance measure can be observed. Here, Jaccard distance yields better results on vectors of stemmed text. Again, the best algorithms are gradient boosting and random forests. With an accuracy of 0.7955 and F1 score of 0.8, gradient boosting and cosine distance on doc2vec

mode	all significant features						all features					
measure	Cosine			Jaccard			Cosine			Jaccard		
V \	C	A	F1	C	A	F1	C	A	F1	C	A	F1
tf-idf U	RF	.7583	.754	GB	.7644	.7624	GB	.7652	.763	GB	.7652	.7626
tf-idf S	GB	.7568	.7533	RF	.7727	.7696	GB	.7773	.7759	GB	.7719	.7701
d2v U	GB	**.8015**	.8053	RF	.7447	.7502	GB	**.7955**	.8	GB	.7341	.7398

Table 5: The best classifiers dependent on distance measures and document vector representations while using all significant features or all features.

vectors generate the best results, which are nearly as high as those achieved when using only significant features in the classification.

In Table 5, results of comparisons of the different document representations and distance measures can be found for the classification on all significant features and the prediction based on all available features.

When using only the 33 features in the classification process, which were found to be significant in [9] or using only the features which were significant in the TrueImpactDataset [11] as well as in SeminalSurveyDBLP, the same pattern occurred for each combination of document representation and distance measure. While cosine distance is better suited in combination with doc2vec, when using tf-idf Jaccard distance produces better results. Gradient boosting performed best for the task of classification based on the 33 significant features. The highest accuracy of 0.7856 with an F1 score of 0.791 was achieved when using cosine distance and doc2vec vectors. Using only the intersection of significant features in both datasets resulted in an accuracy of 0.7871 with F1 score of 0.7929 which is marginally better. It is reasonable to observe a higher accuracy in the classification task when observing less but simultaneously more relevant features for a dataset. Focusing on a subset of meaningful features can lead to better generalization of the resulting model [3] and thus higher accuracy after cross-validation compared to usage of all, partially irrelevant features.

Table 6 holds detailed values for the publication vectors and distance measures for all features significant for Herrmannova et al. [9] and for the features which were significant in their dataset and in the SeminalSurveyDBLP dataset.

5.4 Discussion

Herrmannova et al.'s best performing algorithm was naïve Bayes, for our dataset gradient boosting and random forests achieved the best results [9]. While they only evaluated tf-idf in combination with cosine distance, we showed the usefulness of Jaccard distance in the context of single feature classification. In multi feature scenarios, classification based on doc2vec document representation achieved the highest scores.

Our best accuracy of 0.9288 was scored when classifying on sole tf-idf document vectors of publications. This outcome suggests our proposed dataset is

mode	ASF in [9]						ASF for us and in [9]					
measure	Cosine			Jaccard			Cosine			Jaccard		
V \	C	A	F1	C	A	F1	C	A	F1	C	A	F1
tf-idf U	GB	.7561	.7519	GB	.7652	.763	GB	.7409	.7353	GB	.7515	.7496
tf-idf S	GB	.7583	.7563	GB	.7652	.7623	GB	.7485	.745	GB	.7606	.7588
d2v U	GB	**.7856**	.791	GB	.7402	.7461	GB	**.7871**	.7929	RF	.7424	.7444

Table 6: The best classifiers dependent on distance measures and document vector representations while using all significant features (ASF) for TrueImpact-Dataset [11] contrasting the usage of features only significant in TrueImpact-Dataset and SeminalSurveyDBLP dataset.

too artificially engineered due to its creation process for solving the classification task on document vectors representing the publications alone. It could be argued that this flaw diminished when using document vectors. When using features extracted from a citation distance network, an accuracy of 0.8015 was achieved with doc2vec and cosine on the combination of all significant features. Single feature prediction lead to an accuracy which was 0.056 worse than the multi feature case so the combination of (significant) features results in better performance in the classification task at hand.

For Herrmannova et al. [9], features in a CDN from groups B and D which represent references as well as ones from group C which represent citations were relevant in the classification. We observed features from groups E and C which represent citations to be performing well for the SeminalSurveyDBLP dataset. This suggests a higher influence of citations compared to references in determining if a publication is seminal or survey. Multiple reasons could lead to this: Most of the referenced papers of a publication are not read by the authors [34]. Another aspect might be that references should not be weighted the same, as they do not contribute equally to a paper [40,23]. Different referencing practices in computer science [32] could also contribute to this finding. In the case of citations and the publication, the different influences might cancel each other out as a seminal paper is probably referenced as its approach is used while surveys might rather summarize a topic.

The results of the single feature prediction using doc2vec are promising for cases, in which a publication was not yet able to accumulate lots of citations. In several cases, features from group D which are already fully known at the time of the release were the most descriptive ones.

6 Conclusion and Future Work

We revaluated the identification of seminal and survey papers based on semantometrics derived from our proposed SeminalSurveyDBLP dataset. We used tf-idf and doc2vec as document vector representation, cosine and Jaccard distance as well as a multiplicity of machine learning algorithms. Using multiple features de-

rived from semantic distances in the citation distance network of a publication is highly useful for the classification in seminal and survey (accuracy 0.8015, F1 score 0.8053). Single feature classification worked well on features from group E regardless of the underlying document vector representation and distance measure. Gradient boosting was repeatedly amongst the best performing machine learning algorithms.

Contrasting Herrmannova et al.'s findings which suggest features of groups B, C and D were good predictors in the classification task [9], our experiments point to features from group E and C as best suited. This oppositional observation could be explained by the nature of the used datasets. While Herrmannova et al.'s dataset is multi disciplinary, SeminalSurveyDBLP is focused on papers from computer science. This area is known to behave differently from other domains as highly cited publications limit their references to a narrow community [32].

A worthwhile extension of the evaluated approach could be the usage of LDA [2] document vectors combined with a suitable (weak) metric such as earth mover's distance or the incorporation of more statistical features such as entropy [8]. Other distance metrics or text representations such as GloVe [25] could also contribute to better results. Automatic feature engineering with deep feature synthesis [12] could produce more descriptive features which in turn might lead to higher accuracy.

Another direction for further efforts could be hyperparameter tuning via grid search or the incorporation of more sophisticated machine learning algorithms such as gpt-2 [27] as classifier.

Lastly, a thorough automatic evaluation of our dataset or even the creation of a manually evaluated dataset with even more publications and full texts spanning multiple research areas would be desirable. As the SeminalSurveyDBLP dataset contains information on years of publications, it can be used to analyse if the classification performance changes for papers which have had different periods of time to accumulate citations. A new dataset which does not concentrate on providing similar distributions of citations and references but instead purely holds publications which received a best paper award as well as surveys could describe another interesting bibliographic perspective to analyse.

References

1. D. W. Aksnes: Characteristics of highly cited papers. In: *Research Evaluation* 12(3): 159–170 (2003).
2. D. M. Blei, A. Y. Ng, and M. I. Jordan: Latent Dirichlet Allocation. In: *Journal of Machine Learning Research* 3: 993-1022 (2003).
3. A. Blum and P. Langley: Selection of Relevant Features and Examples in Machine Learning. In: *Artif. Intell.* 97(1-2): 245-271 (1997).
4. http://www.core.edu.au/.
5. B. Cronin and L. I. Meho: Using the h-index to rank influential information scientists. In: *JASIST* 57(9): 1275-1278 (2006).
6. E. Garfield: Can Citation Indexing Be Automated? In: *Essays of an Information Scientist* 1: 84-90 (1964).

7. S. Gerrish and D. M. Blei: A Language-based Approach to Measuring Scholarly Impact. *ICML 2010*: 375-382.
8. R. J. Gillies, P. E. Kinahan and H. Hricak: Radiomics: Images Are More than Pictures, They Are Data. In: *Radiology* 278(2): 563-577 (2016).
9. D. Herrmannova, P. Knoth, and R. M. Patton: Analyzing Citation-Distance Networks for Evaluating Publication Impact. *LREC 2018*.
10. D. Herrmannova, Petr Knoth: Semantometrics in Coauthorship Networks: Fulltext-based Approach for Analysing Patterns of Research Collaboration. In: *D-Lib Mag.* 21(11/12) (2015).
11. D. Herrmannova, R. M. Patton, P. Knoth, and C. G.Stahl: Citations and readership are poor indicators of research excellence: Introducing TrueImpactDataset, a New Dataset for Validating Research Evaluation Metrics. In: *Proceedings of the 1st Workshop on Scholarly Web Mining, 2017*.
12. J. M. Kanter and K. Veeramachaneni: Deep feature synthesis: Towards automating data science endeavors. *DSAA 2015*: 1-10.
13. P. Knoth, D. Herrmannova: Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. In: *D-Lib Magazine* 20(11/12) (2014).
14. F.-T. Krell: The poverty of citation databases: data mining is crucial for fair metrical evaluation of research performance. In: *BioScience* 59(1): 6-7 (2009).
15. V. Kumar et al.: Radiomics: the process and the challenges. In: *Magnetic Resonance Imaging* 30(9): 1234-1248 (2012).
16. Q. V. Le and T. Mikolov: Distributed Representations of Sentences and Documents. *ICML 2014*: 1188-1196.
17. M. Ley: DBLP - Some Lessons Learned. In: *PVLDB* 2(2): 1493-1500 (2009).
18. A. Livne, E. Adar, J. Teevan, and S. Dumais: Predicting citation counts using text and graph mining. *iConference 2013 Workshop on Computational Scientometrics*.
19. M. H. MacRoberts and B. R. MacRoberts: Problems of citation analysis: A study of uncited and seldom-cited influences. In: *JASIST* 61(1): 1-12 (2010).
20. H. F. Moed: The impact-factors debate: the ISI's uses and limits. In: *Nature* 415: 731–732 (2002).
21. S. L. Montolio, D. Dominguez-Sal, and J.-L. Larriba-Pey: Research endogamy as an indicator of conference quality. In: *SIGMOD Record* 42(2): 11-16 (2013).
22. R. M. Patton, D. Herrmannova, C. G. Stahl, J. C. Wells, and T. E. Potok: Audience Based View of Publication Impact. *WOSP@JCDL 2017*: 64-68.
23. R. M. Patton, C. G. Stahl, and J. C. Wells: Measuring Scientific Impact Beyond Citation Counts. In: *D-Lib Magazine* 22(9/10) (2016).
24. Pedregosa et al.: Scikit-learn: Machine Learning in Python. *JMLR 12*: 2825-2830 (2011).
25. J. Pennington, R. Socher, and C. D. Manning: GloVe: Global Vectors for Word Representation. *EMNLP 2014*: 1532-1543.
26. D. Pride and P. Knoth: Incidental or Influential? - Challenges in Automatically Detecting Citation Importance Using Publication Full Texts. *TPDL 2017*: 572-578.
27. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever: Language Models are Unsupervised Multitask Learners. (2019).
28. L. M. A. Rocha and M. M. Moro: Research Contribution as a Measure of Influence. In: *Proceedings of the 2016 International Conference on Management of Data*: 2259–2260 (2016).
29. M. Schreiber: The influence of self-citation corrections on Egghe's g index. In: *Scientometrics* 76(1): 187-200 (2008).

30. P. O. Seglen: Why the impact factor of journals should not be used for evaluating research. In: *British Medical Journal* 314(7079): 498-502 (1997).
31. https://doi.org/10.5281/zenodo.3258164
32. X. Shi, J. Leskovec, and D. A. McFarland: Citing for high impact. *JCDL 2010*: 49-58.
33. T. H. P. Silva, M. M. Moro, A. P. C. da Silva, W. Meira Jr., and A. H. F. Laender: Community-based endogamy as an influence indicator. *JCDL 2014*: 67-76.
34. M. V. Simkin and V. P. Roychowdhury: Read Before You Cite! In: *Complex Systems* 14(3) (2003).
35. M. V. Simkin and V. P. Roychowdhury: Copied citations create renowned papers? In: *Annals of Improbable Research* 11(1): 24–27 (2005).
36. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su: ArnetMiner: Extraction and Mining of Academic Social Networks. *KDD 2008*: 990-998.
37. M. Valenzuela, V. Ha, and O. Etzioni: Identifying meaningful citations. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).
38. A. D. Wade, K. Wang, Y. Sun, and A. Gulli: WSDM Cup 2016: Entity Ranking Challenge. *WSDM 2016*: 593-594.
39. R. Whalen, Y. Huang, A. Sawant, B. Uzzi, and N. Contractor: Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science. *ASCW'15 Workshop at Web Science 2015*: 6-8.
40. X. Zhu, P. D. Turney, D. Lemire, and A. Vellino: Measuring academic influence: Not all citations are equal. In: *JASIST* 66(2): 408-427 (2015).